# Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients

**Agnes Gruenerbl**
Embedded Intelligence, DFKI
Kaiserslautern, Germany
first.last@dfki.de

**Venet Osmani**
CREATE-Net
Trento, Italy
first.last@create-net.org

**Gernot Bahle**
Embedded Intelligence, DFKI
Kaiserslautern, Germany
first.last@dfki.de

**Jose C. Carrasco**
CREATE-Net
Trento, Italy
first.last@create-net.org

**Stefan Oehler**
Tilak
Hall in Tirol, Austria
first.last@umit.at

**Oscar Mayora**
CREATE-Net
Trento, Italy
first.last@create-net.org

**Christian Haring**
Tilak
Hall in Tirol, Austria
first.last@tilak.at

**Paul Lukowicz**
Embedded Intelligence, DFKI
Kaiserslautern, Germany
first.last@dfki.de

## ABSTRACT

In this paper we demonstrate how smart phone sensors, specifically inertial sensors and GPS traces, can be used as an objective "measurement device" for aiding psychiatric diagnosis. In a trial with 12 bipolar disorder patients conducted over a total (summed over all patients) of over 1000 days (on average 12 weeks per patient) we have achieved state change detection with a precision/recall of 96%/94% and state recognition accuracy of 80%. The paper describes the data collection, which was conducted as a medical trial in a real life every day environment in a rural area, outlines the recognition methods, and discusses the results.

### Author Keywords

smart phone; bipolar disorder; real-life study; state recognition; state change detection

## INTRODUCTION

Cognitive, mental and emotional disorders are an obvious application field for activity recognition. The symptoms of such diseases manifest themselves in behavior changes so that activity aware systems can be used as core diagnostic instruments. The value of such instruments is amplified by the fact that psychiatrists currently have few objective and reliable alternatives. Whereas a physician attending a broken leg can make an X-ray to see exactly what he is dealing with, most of the time psychiatrists have to rely on a patient's subjective recollection of his or her behavior. The problem is even more significant for the patients themselves. In contrast to e.g. a hypertension patient, who only needs to regularly measure his blood pressure and compare the results to thresholds given to him by his doctor, a person with a cognitive or mental disorder has no such simple instrument. The closest thing to a "measurement" are self-assessment questionnaires that can be time consuming and only rely on subjective recollections and the patients' self-perception. As a consequence patients often end up visiting the doctor very late, which makes treatment more difficult and often leads to prolonged hospitalization.

While the benefit of an "objective measurement" based on activity recognition is clear, developing and implementing such a system is difficult for many reasons. First, having mental patients wear complex sensors on a daily basis is often not practicable. Second, since there are no reliable diagnostic instruments, getting enough ground truth for training and testing involves a huge effort in terms of long running trials. Finally, the fact that behavior can vary strongly on a daily basis, independently of illness-based effects, makes recognition difficult. As a consequence, very little work exists on diagnostic work using pervasive sensors in real world environments. Overcoming such difficulties, in this paper we demonstrate how smart phone sensors, specifically inertial sensors and GPS traces, can be used as an objective "measurement device" for aiding psychiatric diagnosis. In a trial with 12 bipolar disorder patients conducted over a total (summed over all patients) of more than 1000 days (on average 12 weeks per patient) we have achieved state change detection with a precision/recall of 96%/94% and state recognition accuracy of 80%.

### Related Work

The usage of wearable and pervasive technology in healthcare has already been explored in numerous publications. Overviews include [12], [2], [21], [16] and [1]. Specific examples range from assisting elderlies with cognitive impairment [18], to monitoring children's developmental progress using augmented toys and activity recognition [22]. In the


AH '14, March 07 - 09 2014, Kobe, Japan

http://dx.doi.org/10.1145/2582051.2582089

area of mental health the majority of systems deployed to date focus on supporting self-monitoring. Systems that require patient feedback through questionnaires or text messages are described in [3] and [23]. Other systems, like [4], [10], and [15] present self-report smartphone applications. Burns et al. [4], for instance, introduce a smart phone application for mood prediction of depressive patients. However, it requires constant interaction and feedback of the patient. LiKamWa et al. [10] display an approach, which - again requiring constant mood input from the user - tries to infer mood through an Iphone application. Furthermore, the "Optimism App" [15] was developed to log self-reported mood, activities and quality of sleep in order to monitor depression. Simpson et al. [19] apply interactive voice response self-monitoring for alcohol abuse disorder patients.

In terms of automatic recognition of mental state much less work exists, in particular work involving real world studies and off the shelf devices like smart phones. In [7], the usage of an indoor location system to assess the state of dementia patients is presented. Massey et al. [13] describe an experimental analysis of a mobile health system for mood disorders where they introduce different possible sensors for mood detection, yet focus on technical aspects like line of sight and reception rate, optimal coverage and optimal placement of on-body sensors. Two publications close to the work presented in this article are the research done by a group from Denmark [6] and the previously mentioned [4] that introduces a mobile phone application which employs machine learning models to try to predict patients' mood (of depressive patients). Here however, the ground-truth is fully self-rated, no objective psychological or psychiatric assessment is performed. In [6], Frost et al. use a self-developed smart phone application to record subjective and objective data from patients suffering from bipolar disorder. Even though their main focus lies on self-reported information they utilize a sidetrack of using coarse objective sensor data (acceleration fragments and phone call statistics) to calculate predictions of simple tendencies of the patient's mental state (state forecast) in order to compare it to the forecast drawn from the self-reporting data. By contrast, our work goes into far more depth in the area of classification, uses location sensors in addition to acceleration and instead of social interaction sensing compares the results to an objective, diagnostic ground-truth on a day to day basis. In previous works our group has also discussed the basic concepts of using smart phones for the management of bipolar disorder [20] and used a smaller (6 patients) data set from a preliminary experiment to detect correlation between selected sensor data and self-reported state ([8] and [17]).

## BIPOLAR DISORDER AND ACTIVITY RECOGNTION

Bipolar Disorder [19] is a common and severe form of mental illness characterized by repeated relapses of mania and depression. Thus, people suffering from the disorder may experience - often in rapid succession - periods of manic, normal and depressive state. The current standard for diagnosis of bipolar disorder uses subjective clinical rating-scales based on self-reporting that were developed in the early 1960s (e.g. HAMD, BRAMS scales) and other more recent variations of them (e.g. BSDS). While the efficacy of these scales has been proven in diagnosing bipolar disorder, they have their drawbacks, as they are a potential source of subjectivity in the diagnosis. Additionally, the diagnosis requires the attendance of a physician. Pharmacotherapy is the main treatment currently offered, but its effectiveness critically depends on the timing of application. Thus, therapy can be very effective if administered at the beginning of a patient's transition to a different state (e.g. from normal to depressive). However, it is much less so when applied only after severe symptoms have persisted for a significant time. As a consequence, a promising form of intervention is teaching patients to recognize and manage early warning signs (EWS). A recent systematic review of this approach found that 11 randomized controlled trials (RCTs) involving 1324 patients show the efficacy of interventions that include EWS self-recognition [11]. However, this involves a very significant training effort (which is difficult to finance) and strongly depends on the patients' compliance and discipline. Thus, it is not always practical or even possible and therefore of limited use.

### The Envisioned Use of Activity Recognition

Following the above considerations and elaborate discussions with the psychiatrists (see also [20]), the aim of this work is to demonstrate that smart phone based activity monitoring can be used as an objective "measurement instrument" that detects state changes in order to ensure that as soon as they occur, appropriate treatment can be administered. It is important to mention that:

1. The recognition results are not meant to automatically trigger medication. There is no danger that a false recognition would trigger potentially dangerous wrong medication.
2. Required "reaction times" are on a time scale of a few days rather than a single day. In fact, radical change seldom happens from one day to the next.

Overall the envisioned usage scenario for the recognition system is to provide daily updates to the doctors and possibly the patients who would then look at the trend evolving on the scale of a few days and, if the trend points towards a negative state change, make sure that an examination is scheduled. This means that for our work:

1. Change detection is more important than the recognition of a particular state.
2. Recognition does not need to be perfect to be useful.

More important than perfect recognition is that the results are achievable in the real world, in a setting not only realistic but actually real. This entails genuine patients and no constraints on where and how to wear the phone, non-technology-savvy users and irregular availability of data from different sensors.

## DATA COLLECTION

### Trials Setup

To collect the required data a medical trial was set up at a psychiatric hospital. The hospital is located in Hall in Tirol, a small rural-area town 15 km east of Innsbruck, the capital city of Tirol, Austria. The trial was set up as a so-called uncontrolled, not randomized, mono-centric, observational study

and was approved by the ethics board of the Innsbruck University Hospital. It aimed at recruiting between 10 and 15 patients for a 12-week participation in the trial. The number of patients and the duration of the trials were limited by the resources available within the project for paying medical and technical support staff. Inclusion criteria were: age between 18 and 65, ability and willingness to deal with current smartphones, being "contractually capable", and a diagnosis of bipolar disorder categorized by ICD-10, F31 (by the International Classification of Diseases and Related Health Problems), with frequently changing episodes. The participation in the study was voluntary and quitting it would not affect the therapy in any way. For each patient the trial proceeded as follows:

1. Patients were recruited during stationary treatment at the clinic by clinic psychiatrists.
2. The trial started with an initial examination. After that the patients were given the phone and data collection began. Note that stationary treatment does not mean a "lock up". Instead the patients would stay in the hospital overnight and attend therapy, but were free to move around the hospital compound and the town close by.
3. Patients were released whenever it was medically advisable. In general, this meant 1 or 2 weeks after the start of the trial, but would come to the hospital for an examination every 3 weeks.
4. A final examination was performed at the end of the trial.

The psychological state examination comprised of 4 standardized scale-tests. The Hamilton Depression Scale (HAMD) and the Common Depression Scale (ADS) were used for determining depression, and the Young Mania Rating Scale (YRMS) and the Mania Self-Rating Scale (MSS) were used for determining mania. These psychological scale-tests were performed by a specifically trained psychologist (clinical psychologist). The examinations resulted in an assessment on a scale between -3 (heavily depressed) and 3 (heavily manic) with intermediate steps of depressed, slightly depressed, normal, slightly manic, and manic. At times the psychologists would even give half grades.

It was agreed to have an examination appointment at the beginning of the study, at the end and only every 3 weeks in-between because more frequent scale-tests would result in a learning effect and thus bias the outcome. To ameliorate the scarcity of ground-truth, well trained and experienced clinical psychologist talked to the patients on the phone in-between measurement appointments. In this way, it was possible to estimate the patients' state and possible changes compared to the preceding scale-test measurement point.

### The Patients

Overall 12 patients participated in the trials between november 2012 and august 2013 (11 female, 1 male, age between 25 and 65). Some patients dropped out early (p0202 and p0602), some (p0502 and p0802) even extended the trial. The evolution of the state of the individual patients during the trial is shown in Figure 1. Note that patients p0202 and p0402 show no change of state during the entire trial period. As a consequence they are not considered in this paper.

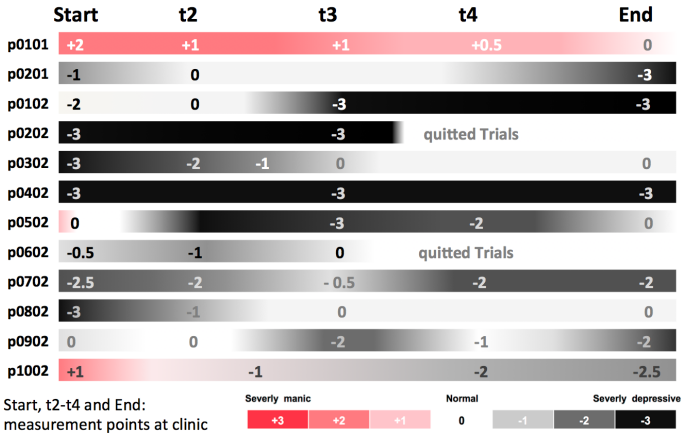


**Figure 1. The evolution of the state of the individual patients during the course of the trial.**

### Data Collection

Each patient was given an Android smart phone running a logging application developed by our group [8]. This application contained two major components: a data logger (using the standard Android API) with the patient having the option of turning off the logging at any time and a self-assessment questionnaire (set to pop-up at the end of the day). After finishing the questionnaire the patient would be asked whether they were comfortable with logging the day's data. If the patient did not agree all data collected during that particular day would be deleted. Otherwise it would be stored on the SD card. This protocol was a pre-condition of the ethics board approval. However during the entire trials there was no case of a patient asking to delete data. The data from the SD card would be copied during the periodic examination and stored in a form that would not reveal the patients identity to the researchers working on the data later on. Clearly in a "productive" system the data would need to be transmitted wirelessly at the end of each day. However, for the purpose of our research the SD card option was more reliable and allowed us to simplify data security issues.

### Ground Truth Extraction

As described in the introduction the aim of this work has been to determine to what degree sensor data can be used to generate an "objective measurement" of a patient's state. This means, that unlike in our earlier study [8], we had to compare the output of our system with the results of the examinations, not with the self-assessment. Thus, there arose the question of how to translate the examinations taking place every three weeks plus some occasional information from the phone interviews into enough labeled days to enable training and testing of the recognition systems. Clearly using only the 5 examination days would not be sufficient. Adding only the results of the phone examinations would also not improve the situation much (max. 12 data points per patient), especially as such additional information was only scarcely available (see

also the next section). After consulting the doctors and building on the experience from the previous small-scale trial [22] the following procedure was applied:

1. As a default the period of 7 days before and 2 days after the examinations (and phone interviews) was considered as ground-truth data. The basic assumption is that state changes are gradual and the probability of a major change within a few days is low. Less days were used after the examination because, in general, visits to the doctor may/should lead to behavior changes (which, incidentally, is the whole point of going to a doctor).

2. The default period then was modified (made smaller or larger) based on the self-assessment questionnaires. The reasoning: according to the psychiatrists, even if self-assessment questionnaires often do not agree with the objective psychiatric rating, they do tend to show a high level of consistency. This means that the same real state leads to a similar self-rating. Thus, if ratings do not vary much during the days before an examination, then a major state change is very unlikely.

3. As a last step of ground-truth extraction we looked at the distribution of days for which ground-truth was available within adjoining classes.

Clearly the above ground-truth is not perfect. However, the fact is that no psychiatric examination ever is. Anyway, it is obviously more "objective" than relying solely on the subjective self-assessment. Finally, note that as is described later we randomly select training and testing data in a cross validation approach to reduce the impact of outliers in the ground-truth.

### Data Amount and Quality

In theory, with 12 patients and 12 (or even more) weeks of trial duration more than enough data ($>$ 1000 days) should be available for our purpose. However, in reality, as we are dealing with a real-life setting, a number of factors influenced the amount of data that actually is suitable for training and testing of recognition systems:

1. Patient Compliance. Since the trial had been conducted under uncontrolled conditions, during normal live, there was no way to make sure that the patients always carried their devices with them. In addition, some patients would even switch off some sensors at certain occasions.

2. Patients state evolution. In general the data collected during the trials was not distributed evenly among the states. In fact most patients had a predominant state and many had "marginal" states for which nearly no data was available.

3. Presence of ground-truth. As described in the previous section, ground-truth was only present for the days around the examinations. In addition, for some patients a number of days had to be discarded due to unstable self-assessment. On top of this, only days where ground-truth and sensor data was available at the same time was usable. Unfortunately this was not the case for all of the ground-truth days.

The effect of the factors described above can be seen in Figure 2. The number of days on which there is any data available

| Patients | Total # of Days | GPS | ACC | Total # of Ground Truth Days (GT) | GPS + GT | ACC + GT | GT+ ACC **OR** GPS |
|---|---|---|---|---|---|---|---|
| p0101 | 97 | 32 | 87 | 84 | 26 (3/20/3) | 71 (32/27/12) | 71 (32/27/12) |
| p0201 | 83 | 78 | 81 | 47 | 36 (10/26) | 38 (12/26) | 38 (12/26) |
| p0102 | 75 | 56 | 61 | 52 | 34 (21/13) | 46 (33/13) | 46 (33/13) |
| p0302 | 90 | 72 | 82 | 70 | 51 (11/40) | 60 (18/42) | 60 (18/42) |
| p0502 | 131 | 60 | 128 | 63 | 28 (5/23) | 58 (14/14/30) | 58 (14/14/30) |
| p0602 | 53 | 44 | 49 | 41 | 31(12/19) | 21 (11/10) | 35 (13/22) |
| p0702 | 76 | 58 | 70 | 53 | 31 (24/7) | 42 (34/8) | 42 (34/8) |
| p0802 | 115 | 71 | 110 | 71 | 37 (7/30) | 62 (16/46) | 62 (16/46) |
| p0902 | 91 | 88 | 88 | 48 | 41 (26/15) | 41 (26/15) | 41 (26/15) |
| p1002 | 67 | 30 | 63 | 47 | 19 (11/9) | 40 (29/11) | 40 (29/11) |

**Figure 2. The amount of sensor data (in days) and Ground Truth days (GT) per patient (in brackets: distribution onto the different classes).**

(first column) is quite close to the amount of time that the patients participated in the trial and varies between 53 and 131. In terms of understanding the usefulness of the data the most important are the last 3 columns showing the number of days on which we have sensor data and ground-truth for each sensing modality and for their combinations. Here the values range from 21 days (p0602, Acc + GPS), to 60-70 days for acceleration (Acc). Even though the data situation is more difficult for the GPS + Acc case (by looking at the distribution of data to the classes (in brackets) some critical cases - only 7 or less samples - can be seen), for most cases the amount of data is reasonable for an evaluation using standard techniques such as a 66/33 percentage-split between training and test data and n-fold cross validation. This is particularly true for the change detection method (see section 6).

## FEATURE COMPUTATION

Based on previous initial experiments [8] and discussions with the medical personnel the following mobility based characteristics were assumed to be relevant:

- Physical motion: Patients with depression tend to move less, move less forcefully and overall slower. The opposite is true for manic patients.

- Travel patterns: Most people have their travel routines dominated by a set of places, which they often visit in a certain temporal pattern. These patterns tend to change in both depressive and manic states (become less frequent or more erratic respectively). In addition depressive people tend to travel and stay outside less often.

Clearly this is to be seen as a statistical average and will be strongly person dependent (some people may move more in depressive state than others do when manic). Also, please note that these are not guesses but have been established in previous work. Based on the above the following features were used:

### Acceleration Features

First, the raw signal was resampled to a fixed sampling rate of 5Hz in order to address variations introduced by the Android phones. Afterwards, signal magnitude was calculated making the following feature extraction invariant to phone orientation (which is unknown). As a final preprocessing step all parts where the signal variance falls under a small threshold were removed and interpreted as time periods where the phone was not on the body (the threshold was experimentally derived by

leaving the phone just lying on a stationary surface). Once this preprocessing was completed, a set of features was calculated using a 10 second sliding window. These were RMS, frequency centroid and frequency fluctuation. The latter two were based on the Fourier transform of the signal. Finally, since labels and location data were calculated on a daily basis, all of the features (in all of the ten seconds windows of each day) were aggregated by computing their mean and variance. These were then used for classification on a daily basis.

#### Location Features

The Android phones recorded GPS traces. For privacy reasons, actual translations into real world locations ("movie theatre", "shopping", etc.) were not possible. The ethics board insisted that the co-ordinates had to be translated into an artificial "anonymized" co-ordinate system that merely showed position relative to the users home. Thus, for our work we used the following set of abstract features:

1. The number of distinct locations visited (applying clustering to the GPS point cloud using a 500m threshold)

2. The number of hours outdoors: the number of hours during which the patient was outside at least once.

3. The average time outdoors per hour: the mean of the time spent outdoors within each of the 24 hours of a day

4. The mean of the times of day spent outdoors (calculated by enumerating all hours and averaging the subset spent outside, which can indicate when a patient was most active - e.g. in the morning or afternoon).

5. The variance of the times of day spent outdoors (similar to 4 yet substituting "variance" for "mean". Both features in conjunction provide a clearer picture of the temporal distribution of patient activity.

6. The number of stays outdoors (connected stays outdoors = a consecutive number of GPS data points within 15 minutes of one another. Timestamps more than 15 minutes apart consequently marked a new segment outside.

7. The percentage of time outside in 24 hours: the sum of the duration of all connected stays, divided by 24 hours.

8. The distance travelled: sum of all distances travelled on any particular day.

All of the above features were calculated on a daily basis.

### STATE RECOGNITION

Using the features described in the previous section we first investigated the recognition of individual states using a standard supervised training approach. The recognition was done on a per patient basis (training and testing on the same patient). As we will show later (see section 6), person independence (in this case, training with one patient and testing on another) is probably not achievable. This stands to reason, as every patient behaves differently.

#### Single Modality Classification

With the features described above in place, we first attempted to apply standard pattern recognition techniques to the data to try to identify which state a patient had been in. As is common with supervised learning methods, we performed a percentage split on our dataset, dividing it into 66% training and 33% test samples. The split was performed randomly. The test-set was resampled to ensure that classes were equally represented. For the actual classification, features were first transformed using a linear discriminant analysis [5]. Afterwards, the Naïve Bayes classifier included in Weka [14] was used to estimate classes for the test-set. Other classifiers were tested (k-nearest neighbor, j48 search tree, conjunctive rule learner), but achieved very similar results. Since the per sample class probability distribution output of the Bayesian classifier was used for the fusion of modalities described in the next section, the results presented for acceleration and location individually were also taken from this classifier. The entire process above was repeated 500 times in a cross validation approach with random test/training splits to eliminate artefacts caused by "lucky" or "unlucky" random selections.

#### Classifier Fusion

The previous step resulted in a list of probabilities for all possible classes, for each day, for each modality (acceleration and location). Combining them yielded a final classification for each day data was available for. The combination process was performed as follows: For every day where there was only one modality available, the most probable class of the associated class probability list was chosen. For every day where both acceleration and location provided class estimates, those estimates were fused using this algorithm:

For each class, the ratio of training data available for acceleration and location compared to all training data was calculated. If modality one provided 10 samples of training and modality two provided 5 samples, the ratio would have been 0.66 for one and 0.33 for two. In order to further penalize little available training data, these coefficients where then input into a sigmoid weighting function: $1/(1 + e^{(} - (coeff - 0.5) * 5))$ Finally, the product of estimated class probabilities and coefficients was calculated for each modality. The two vectors of class estimates were then summed up and the highest rated class picked as winner. The above scheme was chosen because available data was often scarce; it is a well-known fact that for supervised training, an adequate amount of training data is required. Thus, it makes sense to trust classifiers more when they are based on a larger amount of data (within reason, of course).

#### Results

The results (accuracy) are summarized in Table 1 below for each patient and each sensing modality plus their fusion. The accuracies for the individual patients are between 66% and 92%, which is reasonable given the application case outlined in section 2. It can be seen that overall the results are best for the location classifier closely followed by the fused classifier. Average precision/recall values are around 80% for location, around 70% for fusion and around 60% for acceleration (see

**Table 1. Percentage of correctly recognized days per patient: GPS only, Acc only and sensor fusion. Absolute number of recognized instances is given in brackets**

| Patients | Fusion | GPS | ACC |
|---|---|---|---|
| p0101 | 70% (70) | 77% (26) | 75% (70) |
| p0102 | 84% (46) | 82% (34) | 76% (46) |
| p0201 | 68% (38) | 77% (36) | 68% (38) |
| p0302 | 82% (60) | 92% (47) | 66% (60) |
| p0502 | 71% (58) | 85% (28) | 72% (58) |
| p0602 | 77% (31) | 71% (31) | 66% (21) |
| p0702 | 74% (42) | 77% (31) | 73% (42) |
| p0802 | 79% (62) | 89% (37) | 77% (62) |
| p0902 | 83% (35) | 85% (35) | 70% (35) |
| p1002 | 68% (43) | 79% (22) | 71% (43) |
| **mean** | **76%** | **81%** | **72%** |

**Table 2. Overall recall and precision for GPS only, accelerometer only and sensor fusion**

| | **Recall** | **Precision** |
|---|---|---|
| LOC | 81,7% | 80,8% |
| ACC | 62,9% | 64,8% |
| **Fusion** | **70,3%** | **74,0%** |

Table 2). A close look at the individual precision/recall values in Table 3 reveals the reason why location performs best. Since there is not enough data the location classifier does not consider medium depression for patient p0502, which is very poorly recognized by the other classifiers. Overall the fused approach has the advantage of considering more data points than either acceleration or location alone since, as described in subsection B (Classifier Fusion) it considers data points covered by either modality. Looking at the individual patients and states and comparing them with the amount of data outlined in Figure 1 (in particular the distribution into classes in brackets) it is clear that many of the poorly faring cases are related to a very small amount of available data.

**Table 3. Precision / recall values for the different states. Most patients experienced 2, some 3 different states during the trials.**

| | **Recall** Fus/Loc/Acc | **Precision** Fus/Loc/Acc | | **Recall** Fus/Loc/Acc | **Precision** Fus/Loc/Acc |
|---|---|---|---|---|---|
| **p0101** | % | % | **p0602** | % | % |
| normal | 84 / 2 / 74 | 75 / 1 / 88 | sl. dep. | 85 / 58 / 57 | 65 / 63 / 69 |
| sl. man. | 55 / 80 / 75 | 65 / 90 / 59 | normal | 73 / 78 / 61 | 89 / 74 / 47 |
| med.man. | 67 / 67 / 80 | 67 / 63 / 79 | | | |
| **p0102** | % | % | **p0702** | % | % |
| depr. | 94 / 84 / 81 | 86 / 86 / 83 | sev.dep. | 91 / 82 / 82 | 80 / 94 / 85 |
| normal | 62 / 77 / 55 | 80 / 73 / 51 | sl. dep. | 0 / 21 / 22 | 0 / 7 / 18 |
| **p0201** | % | % | **p0802** | % | % |
| depr. | 25 / 75 / 40 | 50 / 48 / 41 | depr. | 44 / 70 / 58 | 64 / 57 / 49 |
| normal | 89 / 81 / 76 | 72 / 86 / 75 | normal | 91 / 93 / 84 | 82 / 96 / 88 |
| **p0302** | % | % | **p0902** | % | % |
| depr. | 39 / 94 / 42 | 100/64 / 32 | depr. | 92 / 83 / 73 | 83 / 82 / 76 |
| normal | 100/92 / 73 | 80 / 99 / 80 | normal | 67 / 70 / 57 | 83 / 71 / 53 |
| **p0502** | % | % | **p1002** | % | % |
| sev.dep. | 50 / 50 / 55 | 58 / 74 / 62 | med.dep. | 86 / 68 / 75 | 74 / 91 / 83 |
| med.dep. | 35 / 0 / 56 | 71 / 0 / 39 | sl. man. | 18 / 87 / 32 | 33 / 57 / 22 |
| normal | 69 / 82 / 81 | 85 / 57 / 87 | | | |

## CHANGE DETECTION

In this section we investigate the detection of a state change without explicit recognition of the new state. The main difference to the approach in the previous section is that instead of a classifier that has a model for each state relevant to the patient, we only build a model of a single "default state". All points falling outside this model are classified as a "change".

This approach is motivated by the following considerations:

1. As discussed in section 2 (Bipolar Disorder), from the application point of view detecting a change of state in order to trigger a visit to the doctor is a key functionality. The exact diagnosis is then done by the doctor anyway.
2. The approach of starting with a single default state has the advantage that a new patient that comes to a doctor can be given a device and as soon as initial data has been collected for his current state the device becomes useful (since a change can be detected). There is no need to wait until data for all relevant states has been collected.
3. For 8 out of 10 patients we are dealing with a two state problem anyway.

Since, in the state change detection case we are explicitly building our own probability density functions (PDFs) for the default state rather than using the output of an out of the box WEKA implementation we also use this case to test a more controlled fusion strategy where the distances to the mean are used as weights together with the number of available training points.

### Single Modality Methode

To evaluate whether this change detection is indeed possible given the data we have collected, we devised the following evaluation:

1. Set up a model for a baseline: As described previously, a number of features have been calculated on both location and acceleration data. Furthermore, for each patient, a number of different classes exists in the recorded period, e.g. samples from a "normal", "lightly manic" and "medium manic" episode. To establish a basis for a certain state of mind, two thirds of samples of a given class were randomly taken as a "training set". A multivariate Gaussian distribution was fitted to this subset.
2. Establish a distance measure and decision boundary: To measure how far a sample is from the model established in (1), the Mahalanobis distance lends itself as a natural choice. For all training samples, the distance to the model was calculated and their mean and standard deviation was determined. Using these two parameters, distances of any sample to the model distribution were normalized by subtracting the mean and then dividing by the standard deviation. If a sample's normalized distance was within a certain threshold, they were marked as "no change" and otherwise as "change". For the threshold, a set of values was tested, resulting in the precision/recall-graph presented in the evaluation section.
3. Evaluate the model: Afterwards, all samples not used for training were used to test the model formed in (1) and (2) by calculating their normalized distance to the model, then checking if they lay in or out of the threshold region established in (2). Based on these estimates, precision and recall were calculated.
4. Perform steps (1) to (3) a thousand times in a cross validation approach with different test/training splits to eliminate

the influence of outliers and offset the artefacts caused by the random selection of training samples. Also, evaluate different thresholds from 0 to 5 in steps of 0.1

### Fusion of Location and Acceleration

Both acceleration and location were evaluated according to this algorithm individually. Afterwards, the results of both modalities were fused, both to improve accuracy and to widen the range of days considered (different modalities may provide data for different, though potentially overlapping, sets of days). For days where there was only data from one modality, that modality was used. When both modalities were available, three different approaches were tested:

1. Logical AND: A state change was assumed if and only if both modalities detected one.
2. Logical OR: A state change was assumed if one or both modalities indicated it.
3. Weighted, at the normalized distance stage: Here, we went back to step (3) of the previous algorithm. Normalized distances were calculated for both modalities, then summed up according to the same sigmoid weighting scheme outlined in the state detection classifier fusion section. Thresholds were combined accordingly. Afterwards, the fused distance was simply tested for being smaller or larger than the new threshold.

### Results

The precision/recall-graphs sweeping over different threshold values for the different fusion approaches and modalities are shown in Figure . It can be seen that the weighted fusion approach is by far the best one reaching an optimal precision/recall value of 96%/94%.

The precision/recall values, reached for each patient in the parameters given by this optimal average point, are given in Table 4 (left). Table 4 (right) contains the values for each state (class averaged over all patients). It can be see that except for the precision for patient p0602 (87%) and recall for light manic (86%) all values are well over 90%. The improvement compared to the state detection can be explained by two factors. First, while only two patients have three states, the results for the state detection for these patients is fairly poor which pulls the overall result down. Second, as is clearly visible in Figure, such excellent results are only reached for the weighted fusion approach, which we have developed specifically for the change detection. For this case, we built explicit PDFs and computed point distances rather than rely on the output of a black box Bayes classifier. This is not surprising since proper weighting is a core aspect of classifier fusion. It should be noted that some improvements in state recognition can possibly be gained by investing the same kind of effort into it that we applied to the change detection in this section. The core aim of this article, however, was change detection, which is the reason our efforts are concentrated here.

### Person (In) Dependence

All of the evaluations described so far in this paper were done in a patient dependent mode. We have justified this by the intuition and previous experience that indicated strong person

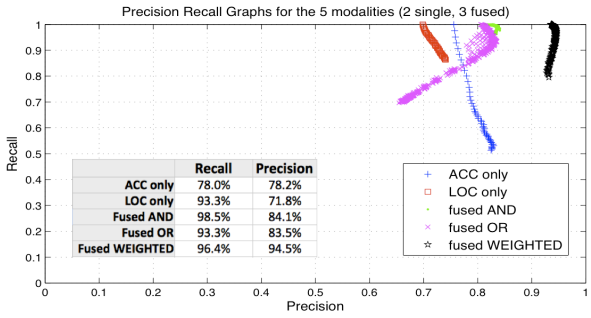


**Figure 3. Precision/Recall graphs for acceleration only, GPS only, fused AND, fused OR and fused WEIGHTED modality.**

**Table 4. Recall and precision values for each patient (right) and each class (left) using the weighted fusion.**

| Patient | Recall | Precision | Class | Recall | Precision |
|---|---|---|---|---|---|
| p0101 | 91.1% | 93.4% | depressive | 99.1% | 95.5% |
| p0102 | 86.2% | 96.8% | heavy depres. | 96.8% | 96.5% |
| p0201 | 97.3% | 92.9% | medi. depres. | 100.0% | 90.6% |
| p0302 | 100.0% | 93.8% | light depres. | 100.0% | 96.8% |
| p0502 | 97.8% | 97.6% | normal | 94.4% | 92.8% |
| p0602 | 100.0% | 87.4% | light manic | 86.7% | 96.7% |
| p0702 | 96.8% | 97.1% | medi. manic | 100.0% | 96.9% |
| p0802 | 95.6% | 95.2% | | | |
| p0902 | 100.0% | 97.1% | | | |
| p1002 | 100.0% | 91.2% | | | |
| **Average** | **96.5%** | **94.2%** | | **96.7%** | **95.1%** |

specific variations in behavior. With the multivariate feature distributions derived in the section "Change Detection" we can back this up with empirical data. To this end, for each of the classes of a patient, the associated distribution was used to calculate the Kullback Leibler divergence [9] to the Gaussian of the same class of all the other patients. While some patients appear to exhibit similar tendencies to one another (patient p0101 and p0302, e.g.), overall divergence values are very high (a mean of 1089). As an example, by comparison, given a 6 dimensional (there were 6 acceleration features) multivariate Gaussian with an expectation value of (10, 10, 10, 10, 10, 10) and the identity matrix as covariance, its KBL divergence to the 6 dimensional normal distribution around 0 is 300. In essence, this confirms what psychologists are fond of pointing out: each patient is an individual and needs to be treated on an individual basis.

## DISCUSSION

The results presented in the previous sections must be seen in the light of a noisy ground-truth and the fact that a patient's behavior cannot be expected to be fully consistent on a daily basis. Even a severely ill person can have a good day. Also, mere change detection (which performed very well in our tests with a precision/recall of 96%/94%) is enough in most cases since the exact diagnosis will be done by the doctor. Furthermore, it would be a potent tool not yet available. In this context we consider our results to be very promising. Clearly, when judging the value of the results presented in this paper one must take into account that for some patients the amount of labeled data and classes was small. This means that it is not possible to say if in large-scale trials we would reach values very close to the 96%/94% for change detection. However, as outlined before, even significantly smaller values

would be sufficient for practical applications. As described in section 2 (Envisioned Use of Activity Recognition) the use case of recognition triggering after persistent occurrence of changed values does not require particularly high accuracies. Even more significant than a few percentage points either way in terms of performance is the fact that the data was collected under conditions that correspond exactly to the way a system would be used in real-life. We have worked with real patients in a rural environment (not necessarily tech savvy) just giving off-the-shelf devices to people with no other supervision than a visit to a doctor every three weeks.


## ACKNOWLEDGMENTS

This work was supported by the MONARCA project (www.monarca-project.eu) from the EU FP7.


## REFERENCES


1. Bardram, J. E. Pervasive healthcare as a scientific discipline. *Methods of information in medicine 47*, 3 (2008), 178–185.
2. Bonato, P. Clinical applications of wearable technology. *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society* (2009), 6580–6583.
3. Bopp, J., Miklowitz, D., Goodwin, G., Stevens, W., Rendell, J., and JR., G. The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study. *Bipolar Disord 12*, 3 (2010), 327–34.
4. Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research 13*, 3 (2011).
5. Fisher, R. A. The use of multiple measurements in taxonomic problems. 179–188.
6. Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., and Bardram, J. E. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *UbiComp*, ACM (2013), 133–142.
7. Gruenerbl, A., Bahle, G., Lukowicz, P., and Hanser, F. Using indoor location to assess the state of dementia patients: Results and experience report from a long term, real world study. In *Intelligent Environments*, IEEE (2011), 32–39.
8. Gruenerbl, A., Bahle, G., Weppner, J., Oleksy, P., Haring, C., and Lukowicz, P. Towards smart phone based monitoring of bipolar disorder. In *Proc. of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*, ACM (2012), 3:1–3:6.
9. Kullback, S., and Leibler, R. On information and sufficiency. 79–86.
10. LiKamWa, R., Liu, Y., Lane, N. D., and Zhong, L. Can your smartphone infer your mood? In *9th ACM Conference on Embedded Networked Sensor Systems (SenSys 2011)*, ACM (Seattle, WA, USA, 2011).
11. Lobban, F. Enhanced relaps prevention for bipolar disorder, 2007. BMC Psychiatry.
12. Lukowicz, P. Wearable computing and artificial intelligence for healthcare applications. *Artificial Intelligence in Medicine 42*, 2 (2008), 95–98.
13. Massey, T., Marfia, G., Potkonjak, M., and Sarrafzadeh, M. Experimental analysis of a mobile health system for mood disorders. *IEEE Transactions on Information Technology in Biomedicine 14*, 2 (2010), 241–247.
14. of Waikato, U. Data mining with open source mashine lerning software. Available at: http://www.cs.waikato.ac.nz/ml/weka/.
15. Optimism. Optimism apps, 2013. Available: www.findingoptimism.com. [Accessed: 09-Jul-2013].
16. Orwat, C., Graefe, A., and Faulwasser, T. Towards pervasive computing in health care - a literature review. *BMC Medical Informatics and Decision Making 8*, 26 (2008).
17. Osmani, V., Maxhuni, A., Gruenerbl, A., Lukowicz, P., C., H., and Mayora, O. Monitoring activity of paitents with bipolar disorder using smart phones. In *Proceedings of MoMM* (2013).
18. Pollack, M. E. The use of ai to assist elders with cognitive impairment for an aging population. *AI Magazine 26*, 2 (2005), 9–24.
19. Simpson, T. L., Kivlahan, D. R., Bush, K. R., and McFall, M. E. Telephone self-monitoring among alcohol use disorder patients in early recovery: a randomized study of feasibility and measurement reactivity. *Drug and alcohol dependence 79*, 2 (2005), 241–50.
20. Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Kappeler-Setz, C., Tröster, G., and Haring, C. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In *Proc. of 2nd Int. Conf. on Pervasive Computing Technologies for Healthcare* (2008), 100–102.
21. Teng, X.-F., Zhang, Y.-T., Poon, C. C. Y., and Bonato, P. Wearable medical systems for p-health, 2008.
22. Westeyn, T. L., Abowd, G. D., Starner, T. E., Johnson, J. M., Presti, P. W., and Weaver, K. A. Monitoring children's developmental progress using augmented toys and activity recognition. *Personal Ubiquitous Comput. 16*, 2 (February 2012), 169–191.
23. Yun, T.-J., Jeong, H. Y., Hill, T. D., Lesnick, B., Brown, R., Abowd, G. D., and Arriaga, R. I. Using sms to provide continuous assessment and improve health outcomes for children with asthma. In *Proc. of the 2nd International Health Informatics Symposium*, ACM (2012), 621–630.